



par Iznogood
<iznogood@iznogood-factory.org>

L'auteur:

Je suis sous GNU/Linux depuis un bon moment et actuellement sur une Debian. Malgré des études électronique, je fais surtout du travail de traduction pour la communauté GNU/Linux.



Résumé:

Ou comment convertir un vieux magazine papier en un document html. Je vais vous expliquer le processus, depuis la numérisation jusqu'à la mise en forme html.

Traduit en Français par:

Iznogood
<iznogood@iznogood-factory.org>

Introduction

J'ai lu que certaines universités américaines aideront ou permettront à Google de numériser leur bibliothèque. Je ne suis pas Google et je suis loin d'avoir une bibliothèque d'université mais je possède quelques vieux numéros d'une revue d'électronique. Et la qualité du papier n'est pas des meilleures: les pages commencent à se détacher, le papier devient gris...

J'ai donc décidé de les numériser car, malgré la disparition de la revue il y a 10 ans, certains (beaucoup) des articles restent actuels! Pour de l'électronique, 10 ans, c'est le moyen âge. C'est dire la qualité des articles.

Matériels

Pour commencer, j'ai eu besoin de rentrer les données dans l'ordinateur. Un scanner est fait pour cela: après avoir vérifié la compatibilité, j'en ai acheté un, un vieux ScanJet 4300C d'occasion mais pas cher et en navigant sur internet, j'ai trouvé les informations de configuration.

Sur la Debian, j'ai installé sane, xsane, gocr et gtk-ocr comme d'habitude avec:

```
apt-get install sane xsane gocr gtk-ocr
```

en tant que root.

Sane et xsane sont des outils de scanner, nécessaires pour le fonctionnement de mon HP.
Gocr et gtk-ocr sont des outils qui permettent de transformer une image en du texte.

Le scanner est un modèle USB:

```
sane-find-scanner
```

me l'a confirmé puis je suis allé dans /etc/sane.d/ pour éditer quelques fichiers:
dans dll.conf, j'ai décommenté

```
hp  
niash
```

et j'ai mis le reste en commentaire.

dans hp.conf et niash.conf, j'ai écrit:

```
/dev/usb/scanner0  
option connect-device
```

et j'ai mis le reste en commentaire.

J'ai modifié le propriétaire du groupe du fichier de périphérique /dev/usb/scanner avec

```
chgrp scanner scanner0
```

et j'ai ajouté iznogood (c'est moi!) comme utilisateur pour me permettre d'utiliser le scanner sans être root:

```
adduser iznogood scanner
```

Un reboot et c'était réglé!

Pour stocker les images, les graveurs de DVD sont suffisamment économiques pour faire le travail, i.e un NEC 3520. J'ai un vieux noyau (2.4.18), le graveur a donc utilisé l'interface SCSI:
Avec modconf, j'ai chargé le module ide-scsi

et j'ai ajouté à /etc/lilo.conf:

```
append="hdb=ide-scsi ignore hdb"
```

puis

```
lilo
```

pour rendre le changement effectif.

Dans /etc/fstab, j'ai ajouté:

```
/dev/sdc0 /dvdrom iso9660 user, noauto 0 0
```

Puis j'ai changé le groupe de sdc0 en cdrom

```
chgrp cdrom sdc0
```

Assez facile.

Logiciel

Pour continuer le processus, j'ai eu besoin de quelques logiciels: sane, xsane, gimp, gocr, gtk-ocr, un éditeur de texte, de html et de l'espace sur le disque dur.

sane est l'interface avec le scanner et xsane, son interface graphique.

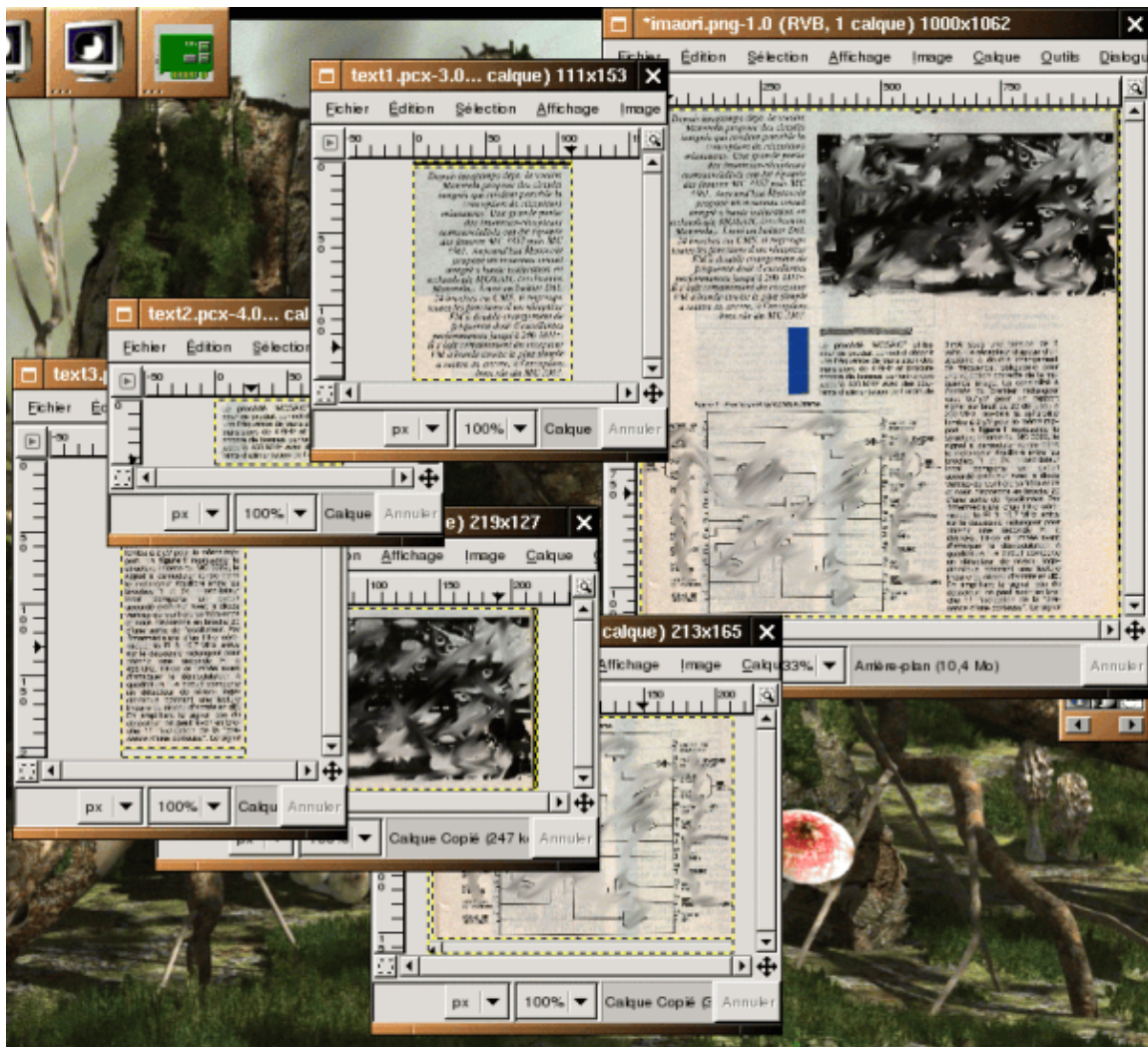
L'idée était de garder une résolution maximale et d'obtenir un fichier de 50 Mo par page, de la stocker sur le disque dur pour la travailler puis la stocker sur un DVDROM.

J'ai mis la résolution à 600 dpi, un peu plus de luminosité et j'ai débuté la conversion. Comme c'est sur une très vieille machine (un PII 350 MHz), cela a pris du temps mais j'ai obtenu une image précise. Je l'ai sauvegardée au format libre png.

Pourquoi une telle résolution et un fichier de 50 Mo? Je voulais garder une résolution maximum pour les archives et pour un traitement numérique ultérieur (et au prix où sont les DVDROM, je n'allais pas me gêner!).

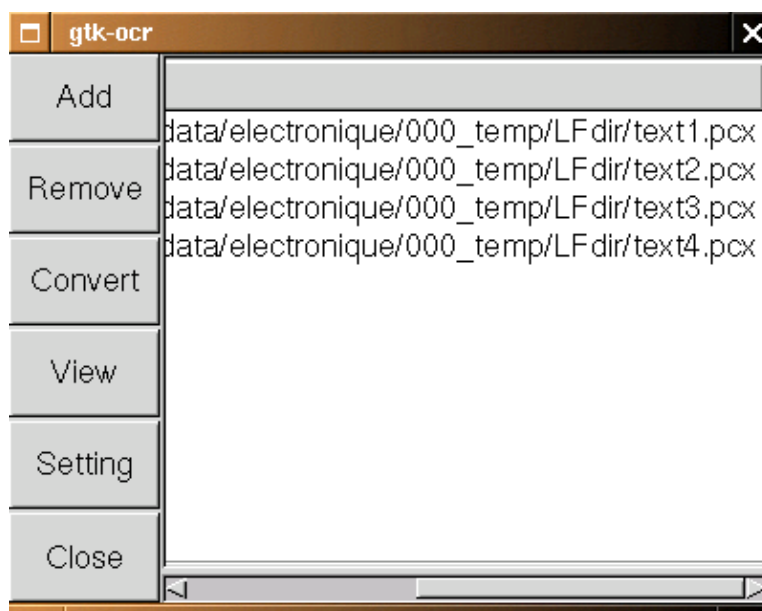
Avec Gimp, j'ai découpé la page en séparant les images graphiques des images textes.

Les graphiques ont été sauvegardés en png avec une taille réduite pour rentrer dans une page html et les images texte n'ont pas été réduites mais changés en niveaux de gris (Outils, Outils des Couleurs, Seuil et Ok puis Image, Mode, Niveau de gris) et sauvegardé avec une extension .pcx pour le traitement avec le logiciel de reconnaissance de caractères.



Vous pouvez voir l'image complète en haut à droite et les parties découpées sur la gauche.
Lorsque vous découpez l'image, vous pouvez enlever les titres car ils prennent trop de place et ne sont pas reconnus par gocr.
J'ai créé un sous-répertoire ima pour les images et les séparer des fichiers .pcx.

C'est là qu'intervient gtk-ocr, l'interface graphique de gocr. gocr est un logiciel de reconnaissance de caractères. Il est très simple à utiliser: j'ai simplement eu à sélectionner les fichiers et gtk-ocr gère tout le reste. Il m'a donné un fichier .txt pour chaque .pcx traité.



Avec un simple


```
cat *.txt > test.txt
```

J'ai eu un test.txt et avec un éditeur de texte, j'ai eu besoin d'effectuer quelques ajustements (suppression de quelques caractères non français, correction de mots...).

Un Copier/Coller dans l'éditeur html, Mozilla composer, dans mon cas et j'ai débuté la composition html (attention de n'avoir que des liens relatifs lorsque vous ajoutez des images).


L'lectronique, c'est bien

Les diodes sont de nos jours peu utilisées isolément. Il est commun de les voir dans des circuits ou de voir leur principe de fonctionnement étendu aux transistors. Néanmoins, on peut encore en trouver tel quel dans des circuits là où il est nécessaire d'installer une voie à sens unique pour le courant. Par exemple, dans un circuit où la polarité est vitale au bon fonctionnement, on peut installer une diode entre les bornes positives et négatives de l'alimentation, qui est passante quand la polarité est mauvaise, créant un court-circuit, détruisant le fusible de protection et, « sauvant » ainsi le reste du montage.



Fabrication

Les diodes sont fabriquées à partir de semiconducteurs et son principe physique de fonctionnement est à la base de tous les composants actifs en électronique.



Scripts Bash

Je me rappelle toujours un prof de maths, lors que j'étais jeune, qui m'avait donné cette maxime:

« Pour être fainéant, il faut être intelligent ».

Bien, je vais commencer par être fainéant !!!! ;-)

Il y a des parties du traitement qui ne sont pas facile à automatiser (création de répertoire, numérisation, découpage avec gimp et création de fichiers). Le reste peut être automatisé.

Il existe un tutoriel en anglais absolument fabuleux sur le scripting en bash, ABS (Advanced Bash Scripting Guide) et j'ai trouvé une traduction française!.

Vous pouvez trouver la version anglaise sur www.tldp.org.

Ce guide m'a permis d'écrire un petit programme. Vous avez le script ici:

```
#!/bin/bash

REPertoire=$(pwd)
cd $REPertoire
mkdir ../ima
mv *.png ../ima/
for i in `ls *`
do
  gocr -f UTF8 -i $i -o $i.txt
done
cd ..
mv ima/ $REPertoire
cd $REPertoire
cat *.txt | sed -e 's/_//g' -e 's/(PICTURE)//g' -e 's/i/i/g' \
-e 's/i/i/g' -e 's/F/r/g' -e 's/î/i/g' > test.txt
```

Le fichier a été rendu exécutable et copié dans /usr/local/bin avec les droits de root sous le nom ocr-rp.

Pour le faire fonctionner, nous devons aller dans le répertoire à traiter et lancer:

```
ocr-rp
```

pwd donnera le chemin du répertoire au script, puis ima est créé à l'extérieur du répertoire et tous les fichiers .png y sont déplacés. Tous les fichiers .txt sont alors listés, traités avec gocr, concaténés dans test.txt et ont subis quelques changements pour avoir des caractères français.

Et nous continuons le même processus qu'avant: Copier/Coller dans Mozilla Composer.

La solution la plus faniéante serait de faire ajouter, par le script, des en-têtes et des bas de page au fichier texte, le sauvegarder et ouvrir directement Mozilla composer mais je suis trop fainéant. Je commencerais demain!!! ;-)

Conclusion

C'était simplement un survol sur les outils permettant la numérisation. Bien entendu, il existe d'autres méthodes, et de bien meilleures, pour le faire. Mais il existe une constante dans le monde GNU/Linux: les outils matériels sont de mieux en mieux supportés chaque année et de plus en plus facile à utiliser. Par exemple, j'ai utilisé un graveur de DVD pour sauvegarder mes images de 50 Mo. L'installation m'a pris 10 minutes et a fonctionné sans problèmes avec k3b (j'ai juste dû faire apt-get install dvdtools dvd+rwtools). Mais cela veut dire qu'avec un vieux PII 350, 192MB RAM, un scanner et un graveur DVD bon marché, un peu d'espace sur le disque dur, vous avez un outil de numérisation suffisamment bon pour « immortaliser » une vieille revue d'électronique.

Voici les liens des outils utilisés pour la numérisation:

- Le scanner est un HP ScanJet 4300C
- sane, www.sane-project.org
- xsane, www.xsane.org
- gimp, www.gimp.org
- gocr, gtk-ocr jocr.sourceforge.net
- ABS est sur www.tldp.org
- et la version française d'ABS est sur abs.traduc.org
- Le graveur de DVD: NEC 3520
- k3b www.k3b.org

<p>Site Web maintenu par l'équipe d'édition LinuxFocus © Iznogood "some rights reserved" see linuxfocus.org/license/ http://www.LinuxFocus.org</p>

<p>Translation information: en --> -- : Iznogood <iznogood/at/iznogood-factory.org> en --> fr: Iznogood <iznogood/at/iznogood-factory.org></p>
--